

What's the value in more moments?

Timothy Daley tdaley@stanford.edu

Department of Statistics and Department of Bioengineering,
Stanford University, California, U.S.A.

Andrew D. Smith andrewds@usc.edu

Department of Molecular and Computational Biology,
University of Southern California, Los Angeles, California, U.S.A.

May 12, 2016

Abstract

The moment-based non-parametric species richness estimator of Chao [8] is one of the most widely used estimators for the number of unobserved species in a sampling experiment. This is due in large part to its simplicity and robustness. This simplicity can also be a drawback, as it only uses a small amount of information contained in the observed experiment, essentially only the first moment. Previous authors, specifically Harris [18] and Chao [7], have presented a general moment-based framework for estimating species richness that includes the Chao estimator. The application of this framework has been stymied by both the lack of deep sampling experiments, where higher moments can be accurately estimated, and the lack of efficient algorithms to properly use this information. Technological advances have filled the former void, allowing for sampling experiments orders of magnitude larger than previously considered. We aim to address the latter by connecting results from the theory of moment spaces and Gaussian quadrature to provide a general moment-based non-parametric estimator of species richness that uses more information through more moments and is computationally efficient. We show this estimator performs well and improves upon the Chao estimator on discrete abundance distributions, the simplest cases of heterogeneity. We demonstrate the performance on a simulated populations taken from emerging high-throughput technologies such as RNA-seq, immune repertoire, and metagenomic sequencing.

1 Introduction

Consider a population composed of an unknown number of distinct species or classes. A sample of individuals is taken with replacement from the population. Once sampled, the species of an individual is known. Following the sampling experiment, the number of sampled individuals from each observed species can be tallied. The problem of estimating the total number of species in the population using information contained in the observed sample is commonly known as the species richness problem. This terminology originated in the context of ecological experiments, but the underlying statistical question is simply to estimate the size of a population through a finite sample.

In theory, the only way to know the total species richness is to sample an infinite number of individuals. Without assumptions on the abundance distribution there can be an infinite number of arbitrarily rare species that a finite sampling experiment cannot possibly capture [18, 5]. Recent work has expanded on this observation, helping to explain the apparent difficulty of constructing practical estimation methods by proving that no unbiased estimator exists in the general case [27,

29]. Even non-parametric maximum likelihood estimators can suffer from extreme instability due to fitting abundance distributions with positive weight on arbitrarily small abundances [38, 19, 30].

The approach introduced by Harris [18] laid the groundwork for practical non-parametric moment-based estimation of species richness that others have built upon [7, 8, 9, 28, 29]. This general strategy relies upon a transformation of the problem of estimating the species richness to the problem of estimating an integral over an unknown distribution given its moments. These moments can be estimated as functions of the observed data and used to form a moment space. This moment space corresponds to the space of species abundance distributions that can explain the observed data. The associated integral has a maximum and minimum value over the constructed moment space, and these in turn correspond to maximum and minimum species richness that can explain the observed data. In agreement with previous results [29] we show that the maximum value is infinite regardless of the observed data. On the other hand the minimum value is finite and in theory will increase towards the true species richness as the dimension of the moment space increases, theoretically providing tighter lower bounds as more information is used.

The problem of estimating an integral over an unknown distribution given its moments is known as Gaussian quadrature, well studied in the numerical integration literature [11]. We use results from the theory of moment spaces to show that the solutions given by Gaussian quadrature are exactly the moment-based estimators of Harris [18]. The simplest case, when the moment space is degenerate, is precisely the Chao estimator [8]. Harris also presented a closed form solution when the moment space is two dimensional, but higher order solutions were previously unavailable. The use of Gaussian quadrature allows us to increase the dimension of the moment space and theoretically use an arbitrary amount of information to construct our estimators. We will show that this is complicated by the ill-conditioning of the procedure. In order to use the maximal amount of information while minimizing the effects of ill-conditioning, we develop an automated method to choose the order of the approximation. This alleviates some of the issues, but to fully avoid some of the severe instabilities we propose bootstrapping the observed histogram to obtain stable estimates.

The outline of the paper is as follows. We introduce the underlying model and the moment-based framework for estimating species richness in section 2. In section 3 we discuss how Gaussian quadrature can solve the moment-based problem. We follow with discussion and analysis of some of the methods for calculating Gaussian quadrature estimates and methods for combatting the inherent ill-conditioning of the problem in section 4. This includes an interesting counter-example specific to our problem to one of the prevailing suggested state of the art methods for numerical quadrature, as discussed in Laurie [23] and Gautschi [13]. In section 5 we compare the performance of the Chao estimator and our moment-based quadrature estimator on discrete abundance distributions. We illustrate the ill-conditioning problem exists even when the order of the approximation is chosen appropriately and how bootstrapping can mitigate the effects of ill-conditioning. Section 6 shows the performance of our estimator on simulated populations taken from large real world high-throughput sampling experiments. We close the paper with a discussion on directions for further investigation.

2 Moment-based estimation of species richness

Suppose that there are S species in the population. Let y_i be the number of observations from species i , the so-called species count of i , for $i = 1, \dots, S$ and let λ_i be the corresponding Poisson rate. Assume that the rates are independently and identically distributed according to some distribution μ with support on the positive real line. The probability that j copies of a random species

are sampled in the experiment is

$$\Pr(y = j) = \int_0^\infty e^{-\lambda} \lambda^j / j! d\mu(\lambda).$$

Note that conditional upon the total number of sampled individuals $N = \sum_{i=1}^S y_i$, the species counts will follow a multinomial distribution with probabilities $p_i = \lambda_i / \sum_{s=1}^S \lambda_s$.

Define the count frequencies to be the number of species observed exactly j times in the sample

$$n_j = \sum_{i=1}^S \mathbb{1}(y_i = j).$$

By the compound Poisson assumption the expected value of n_j is

$$\mathbb{E}(n_j) = S \int_0^\infty e^{-\lambda} \lambda^j / j! d\mu(\lambda).$$

We assume no ordering of the species so that the observed data can be summarized by the count frequencies n_1, n_2, \dots and are a sufficient statistic regardless of S and μ .

The number of unobserved classes, n_0 , is unknown and has expectation

$$\mathbb{E}(n_0) = S \int_0^\infty e^{-\lambda} d\mu(\lambda).$$

Since $S = N + n_0$, any estimate for n_0 immediately provides an estimate for S . Our objective is simply to estimate n_0 , the unobserved species richness, using the observed data.

2.1 Estimating the unobserved species richness

Following ideas introduced by Harris [18], Chao [8] defined the measure ν such that $d\nu(x) \propto x e^{-x} d\mu(x)$. Note that this is a bijection with $d\mu(x) \propto x^{-1} e^x d\nu(x)$ and that both measures have the same support by definition. The moments of ν can be expressed as

$$\begin{aligned} \nu_m &= \int_0^\infty x^m d\nu(x) = \frac{S \int_0^\infty \lambda^{m+1} e^{-\lambda} d\mu(\lambda)}{S \int_0^\infty \lambda e^{-\lambda} d\mu(\lambda)} \\ &= \frac{(m+1)! \mathbb{E}(n_{m+1})}{\mathbb{E}(n_1)}. \end{aligned} \tag{1}$$

Since the observed count frequencies are always unbiased estimates of their expected values, we can substitute them in equation (1) to obtain estimated moments for ν .

If we express $\mathbb{E}(n_0)$ in terms of $d\nu$, then

$$\begin{aligned} \mathbb{E}(n_0) &= S \int_0^\infty x e^{-x} d\mu(x) \left(\frac{\int_0^\infty x^{-1} x e^{-x} d\mu(x)}{\int_0^\infty x e^{-x} d\mu(x)} \right) \\ &= \mathbb{E}(n_1) \int_0^\infty x^{-1} d\nu(x). \end{aligned} \tag{2}$$

Since the S term in $\mathbb{E}(n_0)$ has been absorbed into $\mathbb{E}(n_1)$ in equation (2) above, estimation of n_0 is free of S and equivalent to estimating the integral on the right hand side of the equation (2), $\int_0^\infty x^{-1} d\nu(x)$.

The information we have on ν is expressed through the estimated moments. Therefore we can formulate our moment-based problem as follows:

$$\begin{aligned} & \text{Estimate } \int_0^\infty x^{-1} d\nu(x) \\ & \text{such that } \hat{\nu}_m = \frac{(m+1)!n_{m+1}}{n_1} \text{ for } m = 0, 1, \dots \end{aligned} \quad (3)$$

Estimates for the above integral, subject to the moment constraints, can be substituted in to equation (2) to obtain moment-based estimates of $\mathbb{E}(n_0)$.

Since no unbiased estimator exists for the species richness in the general compound Poisson model [29], there is no estimator of ν that gives unbiased estimates of the integral $\int_0^\infty x^{-1} d\nu(x)$ in equation (3). Therefore we will focus on upper and lower bounds, i.e. extremal estimates, of the integral subject to the moment constraints. These extremal estimates will correspond to lower and upper bounds for the species richness of populations that can explain the observed count frequencies, as reflected through the estimated moments.

2.2 Moment spaces and moment-based estimation

Consider the moment space of all distributions that satisfy the moment constraints

$$\mathcal{V}_M = \{\eta : \eta_m = \hat{\nu}_m \text{ for } m = 0, 1, \dots, M\}.$$

The growth of the moments is bounded, i.e. $\hat{\nu}_m \leq C(m+1)!$ for some constant C . Therefore \mathcal{V}_M converges to a single point as $M \rightarrow \infty$ [34, Proposition 1.5] and the measure ν is uniquely determined by its infinite sequence of moments. This in turn implies the identifiability of the species richness in the Poisson model, assuming one can accurately estimate all moments.

If dimension of the moment space is finite, i.e. $M < \infty$, as it will be in practice, then the space \mathcal{V}_M is convex, closed, and bounded [21]. The linear functional

$$\mathcal{L} : \nu \rightarrow \int_0^\infty x^{-1} d\nu(x)$$

will therefore attain its extrema in the space \mathcal{V}_M . If \mathcal{V}_M is full rank, then the extrema will be achieved at unique discrete distributions of minimal degree contained in the moment space [18]. We can compute the extrema by finding the unique set of points $0 \leq x_1 < \dots < x_P$ and positive weights $\{w_1, \dots, w_P\}$ that satisfy the moment constraints given by

$$\begin{aligned} w_1 + \dots + w_P &= \hat{\nu}_0 \\ w_1 x_1 + \dots + w_P x_P &= \hat{\nu}_1 \\ &\vdots \\ w_1 x_1^M + \dots + w_P x_P^M &= \hat{\nu}_M. \end{aligned} \quad (4)$$

Lower extremal estimates can be obtained by taking M odd and $P = (M+1)/2$. Upper extremal estimates can be obtained by taking M even, $P = M/2$, and fixing $x_1 = 0$. Given the points and weights that satisfy the system of equations (4), the estimated species richness is given by

$$\mathbb{E}(n_1) \int_0^\infty x^{-1} d\nu(x) \approx n_1 \left(\frac{w_1}{x_1} + \dots + \frac{w_P}{x_P} \right). \quad (5)$$

In the case of the upper extremal estimates, the estimated number of unobserved species is infinite, agreeing with the intuition that we can not exclude the possibility of an infinite number of species.

If $P = 1$ and $M = 1$, then the system of equations is simply given by the two equations $w_1 = 1$ and $x_1 = \hat{\nu}_1 = 2n_2/n_1$. The corresponding estimate is $\hat{n}_0 = n_1/x_1 = n_1^2/(2n_2)$, exactly the estimator of Chao [8]. Similarly, Harris [18] derived an analytic solution to the system of equations when 2 points and 3 moments are used. Higher order analytic solutions are impossible as they involve finding the roots of degree 5 or higher polynomials.

If an estimator uses only the first few moments, and equivalently only the first few count frequencies, the estimator risks discarding a significant amount of the available information from the observed sampling experiment. In the case of highly complex populations, the underlying distributions may not be accurately explained by the first few moments. In modern data analysis the sampling experiments can be very large, allowing for accurate estimation of higher moments. We hypothesized that, conditional upon having accurate estimates for the higher moments, using more moments will lead to more accurate estimates of the species richness. This requires solving the system of equations (4). Since analytic solutions are impossible when higher moments are involved, we turn our attention to numerical solutions.

3 Numerical calculation of the extremal distributions with Gaussian quadrature

The system of equations (4) is not unique to the problem of finding extremal values of an integral under moment constraints. We noted that the system is identical to the system of equations that Gaussian quadrature must satisfy [17]. Gaussian quadrature is a well studied numerical integration technique to approximate an integral over a unknown measure given its moments. This suggests an approach to use a larger number of moments by leveraging Gaussian quadrature techniques to solve the problem (3) numerically rather than analytically.

3.1 Gaussian quadrature

A numerical approximation, or quadrature, to the integral $\int_0^\infty f(x)d\nu(x)$ is called Gaussian if it is a discrete approximation, i.e.

$$\int_0^\infty f(x)d\nu(x) \approx \sum_{p=1}^P w_p f(x_p)$$

for some fixed integer P , and the approximation is exact for polynomials of degree $2P - 1$ or less. The points x_1, \dots, x_P and weights w_1, \dots, w_P are collectively called the Gaussian quadrature rules.

Consider the standard basis polynomials: $1, x, \dots, x^{2P-1}$. Gaussian quadrature requires that

$$\int_0^\infty x^m d\nu(x) = \nu_m = \sum_{p=1}^P x_p^m w_p \text{ for all } m = 0, 1, \dots, 2P - 1.$$

This is exactly the system of equations (4) with $M = 2P - 1$. Note that any polynomial can be written as a finite sum of the standard basis polynomials. Therefore, if the system of equations (4) is satisfied then the quadrature rule will be exact for any polynomial of degree $2P - 1$, and vice versa. Furthermore the Gaussian quadrature rules are unique [22]. This implies that the solution to the system of equations (4) is unique for fixed P , and the Gaussian quadrature rules and solutions to the system of equations given by Harris will correspond exactly.

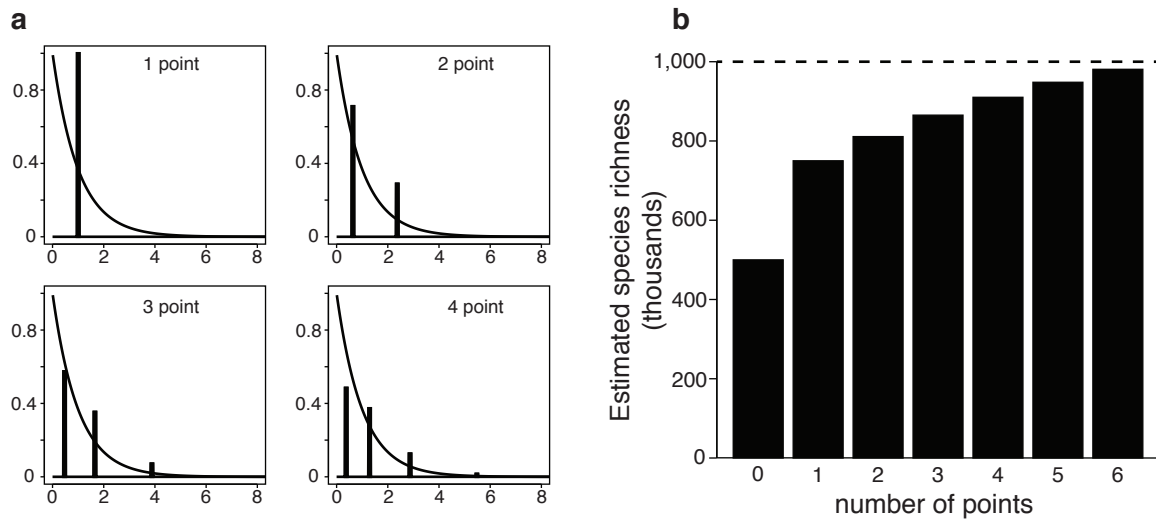


Figure 1: Discrete approximations to a measure by Gaussian quadrature. The compounding distribution $d\mu(x)$ is a Gamma distribution with shape parameter 1 and scale parameter 1. Class counts follow a negative binomial distribution with success probability $p = 1/2$ and size parameter $r = 1$. (a) 1, 2, 3, and 4 point discrete approximations determined by the Gaussian quadrature rules. (b) Quadrature estimates for the number of classes when using the expected counts from one million classes with counts that follow a negative binomial distribution with $p = 1/2$ and $r = 1$.

In the interpretation of Harris [18] and Mao & Lindsay [29] we are approximating ν with a discrete distribution based on the observed moments. Since the map from μ to ν is a bijection, we can interpret the quadrature rules as a moment-based discrete approximation to μ with increasing accuracy as the number of points increases (Figure 1 a). The corresponding estimates for the species richness will in theory converge from below to the true species richness (Figure 1 b).

3.2 Choosing the order of the approximation

Choosing the order, or the number of points, in the quadrature estimator is similar to the problem of determining the order of a moment-based discrete approximation to a distribution. Of critical importance are the moment Hankel matrices,

$$H_P = (\hat{\nu}_{i+j-2})_{1 \leq i, j \leq P} = \begin{pmatrix} 1 & \hat{\nu}_1 & \dots & \hat{\nu}_{P-1} \\ \hat{\nu}_1 & \hat{\nu}_2 & \dots & \hat{\nu}_P \\ \vdots & & \ddots & \vdots \\ \hat{\nu}_{P-1} & \hat{\nu}_P & \dots & \hat{\nu}_{2P-2} \end{pmatrix}. \quad (6)$$

In previous applications [26, 10], the observed data is typically considered to be independently and identically distributed according to the unknown distribution. In this case the moment Hankel matrices are constructed from the usual moment estimators, e.g. for i.i.d. observations x_1, \dots, x_n from an unknown distribution η , $\hat{\eta}_m = \frac{1}{n} \sum_{i=1}^n x_i^m$. This guarantees that the matrices will be non-negative definite. The matrices are only singular when the support of the distribution is discrete with a support size smaller than the dimension of the Hankel matrices. One can then utilize the structure afforded by i.i.d. observations to construct statistical tests of non-negative definiteness and determine the order of the approximation [24, 25].

In our problem, the moments are estimated from the count frequencies and not the usual moment estimators. This means that the moment Hankel matrix will not necessarily be positive semidefinite and we can not utilize the aforementioned tests. To ensure that the estimated quadrature rules correspond to a distribution on the positive real line, we require the estimated quadrature rules to be positive. This can be achieved by ensuring that all Hankel matrices up to order $P + 1$ are positive definite [29]. This requires the first $2P + 1$ moments to be positive, while the quadrature rules only have $2P - 1$ degrees of freedom. To relax the requirement on the number of moments, we first define the shifted moment Hankel matrix $\tilde{H}_P = (\hat{\nu}_{i+j-1})_{1 \leq i, j \leq P}$. We next note that the quadrature rules will be strictly positive if and only if the generalized eigenvalues of the moment Hankel matrix and the shifted moment Hankel matrix are strictly positive [1]. In other words, all solutions z to the equation $\det(H_P - z\tilde{H}_P) = 0$ are positive. A sufficient condition for this is when both H_P and \tilde{H}_P are positive definite [32]. To determine the number of points in the Gaussian quadrature estimator we test the determinants of H_P and \tilde{H}_P for successive values of P , stopping when either is not positive. Positive definiteness will be guaranteed when all lower order matrices have positive determinant by Sylvester's criterion. This requires the first $2P$ moments to be positive and allows us to use more information. In the worst case we can default to the single point case, exactly Chao's estimator.

4 Calculating Gaussian quadrature rules

To naively calculate the Gaussian quadrature rules one can use non-linear numerical solvers to calculate solutions to the system of equations (4) in a brute force manner. Numerical solvers have difficulty in converging in comparison to state of the art quadrature methods. These methods rely upon properties of the underlying distribution, particularly the associated orthogonal polynomials and their relationship with Gaussian quadrature rules [23].

Associated with a measure ν supported on the positive real line is a system of monic orthogonal polynomials $\{\pi_i(x)\}_{i \geq 0}$. They are orthogonal in the sense that $\int_0^\infty \pi_i(x)\pi_j(x)d\nu(x) = 0$ for $i \neq j$. These polynomials satisfy the three term recurrence

$$\pi_{i+1}(x) = (x - \alpha_i)\pi_i(x) - \beta_i\pi_{i-1}(x), \quad \pi_{-1}(x) = 0, \quad \pi_0(x) = 1 \quad (7)$$

for coefficients $\alpha_i, \beta_i > 0$ and $i = 0, 1, \dots$. Since $\pi_{-1}(x) = 0$, the value of β_0 is arbitrary but commonly set to 1. If the measure is uniquely determined by its moments then the three term recurrence is unique and completely characterizes the monic orthogonal polynomials [13].

We will show that the coefficients of the three term recurrence are intimately related to Gaussian quadrature rules so that estimation of the Gaussian quadrature rules can be broken into two parts: (1) estimating the coefficients of the three term recurrence for the moments and (2) estimating the quadrature rules from the estimated three term recurrence.

The mapping from the moments to the quadrature rules is known to be highly ill-conditioned [13]. Small changes in the moments can lead to extreme changes in the corresponding quadrature points and weights. This presents challenges for our approach, as variability in moment estimates can be substantial for higher moments. Most of the ill-conditioning arises from the first part above, while the second part is very well-conditioned [12].

Given the estimated coefficients of the three term recurrence we can construct the truncated Jacobi matrix, denoted J_P . This is a symmetric tridiagonal matrix with $\{\alpha_0, \alpha_1, \dots, \alpha_{P-1}\}$ along the diagonal and $\{\sqrt{\beta_1}, \dots, \sqrt{\beta_{P-1}}\}$ on the main off-diagonals. The quadrature points x_1, \dots, x_P are the eigenvalues of J_P and the quadrature weights w_1, \dots, w_P are the square of the first entry of eigenvectors. The eigenvalues and eigenvectors can be calculated simultaneously by the QR

algorithm. Since J_P is already tridiagonal, the usual Householder transformations are not required. Furthermore the full eigenvector is not needed. This allows for a modified QR algorithm that updates of the diagonal and off-diagonal elements of J_P along with the first entry of the eigenvector at each iteration, without storing the full vectors, giving an algorithm with a time and space complexity linear in the number of moments used [14]. Since the truncated Jacobi matrix is symmetric, this procedure is perfectly conditioned.

Given the moments, the coefficients of the three term recurrence can be calculated by a number of algorithms. The first modern method was based on the Cholesky decomposition of the moment Hankel matrix [14]. The time complexity of this method is cubic in the number of moments. Due to exponential growth in the absolute condition number of the moment Hankel matrix [37], this method can suffer from ill-conditioning when using a large number of moments. To improve the conditioning a recursive method with quadratic time complexity was derived based on Chebyshev's work on orthogonal polynomials [36, 33]. This method requires a known input measure, hopefully close to the unknown distribution, to modify the estimated moments. Accordingly this method is known as the modified Chebyshev algorithm. When no modifying measure is given we shall refer to it as the unmodified Chebyshev algorithm.

It is commonly suggested to use the modified Chebyshev algorithm with a known modifying measure and three term recurrence close to the unknown measure [2, 23]. This is because if the modifying measure is serendipitously chosen to be exactly equal to the unknown measure then it will be recovered with no error. In previously considered applications the moments are assumed to be perfectly observed. The modified Chebyshev algorithm has not been investigated when the moments are not known exactly but instead are estimated with noise.

To evaluate the choice of the modifying measure we will establish a case where the coefficients of the three term recurrence can be derived analytically. We can then examine the situation where the true underlying measure is used as the modifying measure but only random estimates of the true moments are available, essentially if the modifying measure is chosen fortuitously.

4.1 The three term recurrence of negative binomial counts

Suppose that $d\mu$ is a Gamma density with shape parameter r and scale parameter $p/(1-p)$. In this case the number of observations from a random class follows a negative binomial distribution with size r and success probability p and $d\nu(x)$ will be proportional to $x^r e^{-x\phi}$ with $\phi = 1/(1-p)$.

We noted the similarity of the above measure to the measure that generates the classical monic associated Laguerre polynomials [35]. This motivates the following relationships to establish the orthogonal polynomials and the corresponding three term recurrence associated with the density $d\nu$. The proofs can be found in the Supplementary Materials.

Proposition 1. Let $L_i^{(r,\phi)}(x)$ be given by

$$L_i^{(r,\phi)}(x) = \sum_{l=0}^i \binom{i+r}{i-l} \frac{i!}{l!} \phi^{l-i} (-1)^{i+l} x^l. \quad (8)$$

Then $\{L_i^{(r,\phi)}(x)\}_{i \geq 0}$ is the unique system of monic polynomials orthogonal to $d\nu(x) \propto x^r e^{-x\phi}$ for $r > 0$ and $\phi > 0$.

Proposition 2. The polynomials $L_i^{(r,\phi)}(x)$ given in equation (8) satisfy the three term recurrence given in equation (7) with $\alpha_i = (2i + 1 + r)/\phi$ and $\beta_i = (i + r)i/\phi^2$.

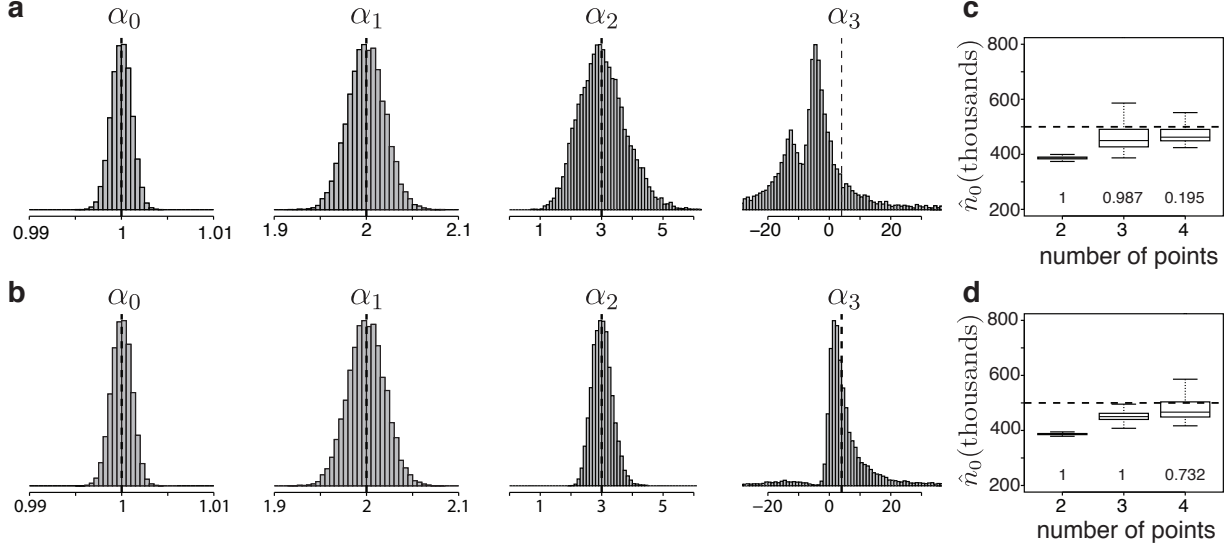


Figure 2: Observed distribution of the estimated coefficients of the three term recurrence and quadrature estimates. For one thousand samples from one million species with counts following independent negative binomial($r = 1, p = 1/2$) distribution, the first four estimated values of α using (a) the modified Chebyshev algorithm with the true underlying measure as the modifying measure and (b) the unmodified Chebyshev algorithm. (c and d) The quadrature estimates and observed fraction of positive quadrature rules corresponding to a and b, respectively. Dashed lines indicate the expected value.

4.2 Estimating the three term recurrence in the presence of error

To investigate the effect that error in estimating the moments has on the stability of the modified and unmodified Chebyshev algorithms, we simulated 1,000 independent samples from a population of one million classes, each class with counts that are independently distributed according to a negative binomial distribution with size parameter $r = 1$ and success probability $p = 1/2$. The three term recurrence was computed using the unmodified Chebyshev algorithm and the modified Chebyshev algorithm, with the modifying measure equal to the true underlying measure ($d\nu \propto xe^{-2x}$) and the true three term recurrence as given in Proposition 2.

The first few estimated coefficients of the three term recurrence (α_0, α_1 , and β_1) were identical for both methods, indicating little effect of ill-conditioning in estimating the first few terms in the three term recurrence (Figure 2). For larger order coefficients, both methods exhibited increasing variance but the unmodified Chebyshev algorithm consistently exhibited less variance. We also observed a left shift in the distribution of the estimated coefficients away from the true values for the modified Chebyshev algorithm that was not present for the unmodified Chebyshev algorithm. This left shift may create problems downstream when estimating quadrature rules with a large number of points. In theory all of the coefficients of the three term recurrence are positive and we make this a strict requirement in our algorithm. The three term recurrence will be truncated and useful information will be thrown away unnecessarily, as exhibited in the lower percentage of positive quadrature rules in Figure 2 c.

The instability we observed in the modified Chebyshev algorithm can only be exacerbated when the true distribution is significantly different from the modifying measure [2]. The instability in calculating the coefficients of the three term recurrence will lead to instability in calculating the

Label	mixture distribution
a	$0.6 \cdot \mathbb{1}(\lambda = 5/13) + 0.4 \cdot \mathbb{1}(\lambda = 25/13)$
b	$0.6 \cdot \mathbb{1}(\lambda = 5/23) + 0.4 \cdot \mathbb{1}(\lambda = 50/23)$
c	$0.7 \cdot \mathbb{1}(\lambda = 5/11) + 0.3 \cdot \mathbb{1}(\lambda = 25/11)$
d	$0.7 \cdot \mathbb{1}(\lambda = 10/37) + 0.3 \cdot \mathbb{1}(\lambda = 100/37)$
e	$0.8 \cdot \mathbb{1}(\lambda = 5/9) + 0.2 \cdot \mathbb{1}(\lambda = 25/9)$
f	$0.8 \cdot \mathbb{1}(\lambda = 5/14) + 0.2 \cdot \mathbb{1}(\lambda = 25/7)$
g	$0.9 \cdot \mathbb{1}(\lambda = 5/7) + 0.1 \cdot \mathbb{1}(\lambda = 25/7)$
h	$0.9 \cdot \mathbb{1}(\lambda = 10/19) + 0.1 \cdot \mathbb{1}(\lambda = 100/19)$

Table 1: Two component distributions.

quadrature rules. Small values of $\vec{\alpha}$ and $\vec{\beta}$ can lead to smaller eigenvalues of the truncated Jacobi matrix, particularly for $\vec{\alpha}$ since the trace of the truncated Jacobi matrix is equal to the sum of the eigenvalues. The small eigenvalues will then be inverted in the estimated integral, leading to large values in the estimated integral and instability in the estimated species richness (Figure 2). We found that in this case the performance of the unmodified Chebyshev algorithm was superior to the modified Chebyshev algorithm, even when the true measure is known, leading to more stable estimates of the species richness (Figure 2 **c** and **d**).

5 Results

The performance of the Chao estimator (CE) has already been benchmarked on uniform populations and shown to perform well in comparison to other estimators [7, Chapter 6]. We examined the performance of the CE in the presence of heterogeneity and how using more information contained in the moments through our moment-based quadrature estimator (MQE) can improve the performance.

5.1 Two component discrete mixture

The simplest case of a perturbed uniform population is when there are two distinct uniform subpopulations. As long as both populations contribute to the first few count frequencies, the MQE should be able to distinguish the one point case from the two point. Examples where the two-point might be indistinguishable from the one point is when few individuals are sampled from one subpopulation, one subpopulation is so prevalent that it contributes little to the first few count frequencies, or individuals in the two subpopulations have similar abundances.

For illustrative examples of two component discrete mixtures, free of the aforementioned problems, we consider populations of one million species with mixtures of 40%, 30%, 20%, and 10% of the population represented five and ten times more abundant than the remaining species, described in Table 1. We refer to the rarer species as the background and the more abundant species as the foreground. We sampled an average of one million individuals from each population one hundred times and calculated the CE and MQE (Figure 3).

In all samples the MQE correctly selected a two point approximation and the estimates were

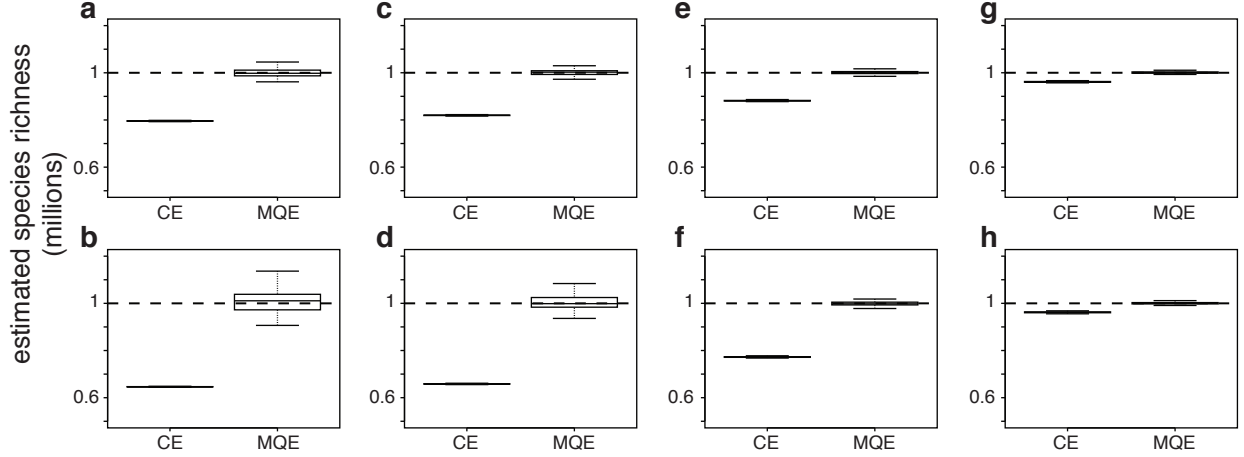


Figure 3: CE and MQE for two component discrete mixtures Chao estimator (CE) and moment-based quadrature estimator (MQE) for populations of one million individuals following a compound Poisson with compounding mixture distributions described in table 1. The population labels **a-h** correspond to the figure label **a-h**.

centered around the true species richness. The CE’s were all lower bounds, as they should be in the presence of heterogeneity [3]. The average improvement of the MQE to the CE ranged from 344,250, or about one third of the population, to 38,443, less than five percent of the population. The difference between the two estimators tended to be less when the foreground population was smaller. In large part, this is because the majority of the foreground population in these cases is observed and contributes little to smaller count frequencies. For example, for population **h**, shown in Figure 3**h**, 97% of the foreground population is expected to be observed. It contributes approximately 3% of the singletons and 9% of the doubletons in expectation but about half of the species in n_3 . Thus the first two count frequencies closely reflect the fact that the vast majority of the unobserved species are uniform and the CE will represent this. In contrast, in population **b** about half the singletons are expected to be from the foreground. The CE has a hard time in estimating the unobserved species due to the distortion from the foreground in the first two count frequencies (Figure 3**b**).

One of the main advantages to the use of CE in estimating species richness is its stability. Even in the ideal equiprobable case the CE performs well compared to maximum likelihood estimators [7, Chapter 6]. The addition of extra information in estimating the species richness through n_3 and n_4 introduces extra variability into the MQE compared to the CE. Obviously there is a bias-variance tradeoff, where estimation using only the first two count frequencies, in the CE, is more biased but less variable while estimation using the first four count frequencies, in the MQE, is less biased but more variable. When the results are measured by root mean square error (RMSE) the MQE shows an order of magnitude lower RMSE than the CE.

5.2 Higher order mixtures: three and four components

The next obvious step is to examine higher order mixtures, specifically three and four component mixtures. We investigated four populations, each composed of one million species. The first two populations had rate distributions that followed a three component mixture,

$$0.6 \cdot \mathbb{1}(\lambda = 5/23) + 0.3 \cdot \mathbb{1}(\lambda = 25/23) + 0.1 \cdot \mathbb{1}(\lambda = 125/23)$$

and

$$0.8 \cdot \mathbb{1}(\lambda = 5/14) + 0.15 \cdot \mathbb{1}(\lambda = 25/14) + 0.05 \cdot \mathbb{1}(\lambda = 125/14),$$

labeled **a** and **b**. The second two populations had rate distributions that followed a four component mixture,

$$0.6 \cdot \mathbb{1}(\lambda = 5/83) + 0.2 \cdot \mathbb{1}(\lambda = 25/83) + 0.1 \cdot \mathbb{1}(\lambda = 125/83) + 0.1 \cdot \mathbb{1}(\lambda = 625/83)$$

and

$$0.8 \cdot \mathbb{1}(\lambda = 5/44) + 0.1 \cdot \mathbb{1}(\lambda = 25/44) + 0.05 \cdot \mathbb{1}(\lambda = 125/44) + 0.05 \cdot \mathbb{1}(\lambda = 625/44),$$

labeled **c** and **d**. The expected number of sampled individuals in all populations is one million, or an average rate of one. For one hundred random samples from each population we computed the CE, MQE, and moment-based two point quadrature estimator (M2QE). For the four component mixture we also computed the moment-based quadrature estimator limited to at most three points (M3QE). The estimates are shown in Figure 4a–d.

In the three component mixtures the MQE chose the correct number of points for the quadrature rules in the vast majority of the samples, 85% from population **a** and 90% from population **b**. In the four component mixtures the MQE chose the correct number of points in far fewer samples, 13% from population **c** and 24% from population **d**. These results indicate that it is difficult to use a large number of points in the quadrature rules as using higher order moments introduce a substantial amount of variability in the estimator. On the other hand, when the true compounding distribution is a discrete mixture the MQE infrequently uses a larger number of points than the true underlying distribution, suggesting that more points should be used if possible.

In all samples the M2QE and MQE were greater than the CE and the MQE was at least as large as the M2QE. Similarly the MQE was at least as large as the M3QE in the four component populations. The improvement of the MQE over the CE in all populations was dramatic, with a median difference of close to 200,000 for the three component populations and over 250,000 for the four component populations. The improvement of the MQE over the M2QE varied between populations. For instance the median difference was 42,000 in population **b**, but is over three times larger in population **a**.

The improvement of the MQE over the lower order estimates naturally brings more variance into the estimates due to the use of more information through more moments. The ill-conditioned nature of the estimation compounds this difficulty, so that some of the MQEs are much, much larger than the true species richness. The effect of this is to typically decrease the performance of the higher order approximations in terms of the root mean square error, though the tradeoff between the number of moments used in the estimates and the RMSE depends on the properties of the population.

To combat the ill-conditioning issue we suggest bootstrapping the observed counts frequencies, which conveniently follow a multinomial distribution [39]. Averaging over the bootstrapped MQE (bMQE) reduces the effect of those estimates that suffer from ill-conditioning and significantly improves the performance of the moment-based quadrature estimator. This results in the bMQE having the highest average species richness estimates and lowest RMSE (Figure 4e–h).

6 Examples

It is always the desire of researchers to test methods on real populations. A few researchers, such as Carothers [6] have designed experiments where the properties of the population are completely

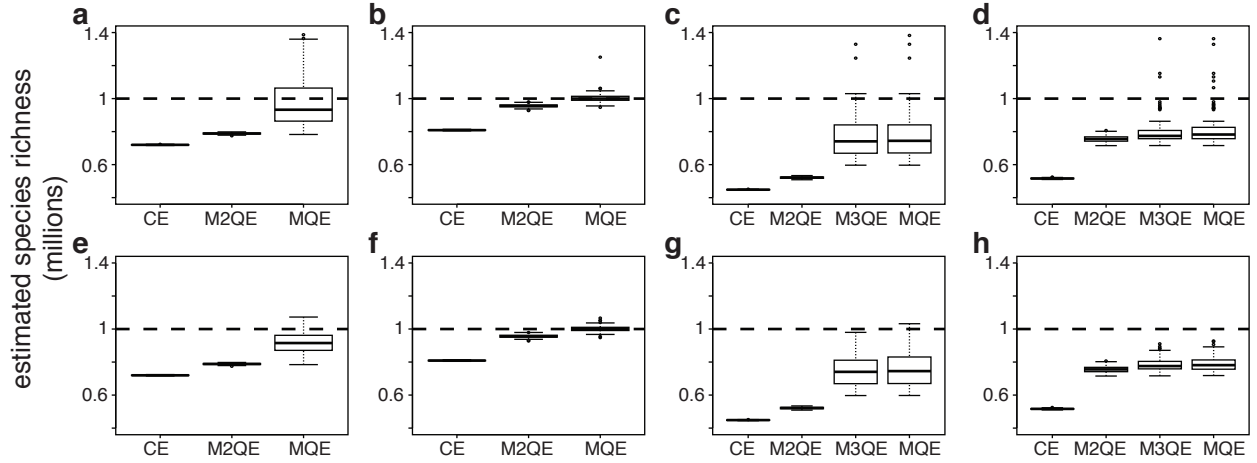


Figure 4: CE and MQE for three and four component discrete mixtures Chao estimator (CE), moment-based quadrature estimator (MQE), and moment-based two point quadrature estimator (M2QE) for populations of one million individuals following a compound Poisson with compounding mixture distributions **a** $0.6 \cdot 1(\lambda = 5/23) + 0.3 \cdot 1(\lambda = 25/23) + 0.1 \cdot 1(\lambda = 125/23)$ and **b** $0.8 \cdot 1(\lambda = 5/14) + 0.15 \cdot 1(\lambda = 25/14) + 0.05 \cdot 1(\lambda = 125/14)$. CE, MQE, M2QE, and M3QE (moment-based three point quadrature estimator) for populations of one million individuals following a compound Poisson distribution with compounding mixture distributions **c** $0.6 \cdot 1(\lambda = 5/83) + 0.2 \cdot 1(\lambda = 25/83) + 0.1 \cdot 1(\lambda = 125/83) + 0.1 \cdot 1(\lambda = 625/83)$ and **d** $0.8 \cdot 1(\lambda = 5/44) + 0.1 \cdot 1(\lambda = 25/44) + 0.05 \cdot 1(\lambda = 125/44) + 0.05 \cdot 1(\lambda = 625/44)$ **e**, **f**, **g**, and **h** are respectively the bootstrapped estimates using the same data as **a**, **b**, **c**, and **d**.

known. Unfortunately such cases are few and small in scale by necessity. Our focus is on large scale sampling experiments where the higher order moments are estimated with more accuracy and contain information that can significantly improve the estimator. In order to simulate large heterogenous populations with known properties and are similar to populations encountered in real world sampling experiments we took a numerical simulation approach. We took large observed experiments as the population and sampled with replacement from the experiment to simulate sampling individuals. We examined several situations where species richness estimators have been applied and used for inference, specifically high-throughput RNA-seq, immune repertoire, and metagenomics.

To simulate high-throughput RNA-seq experiments we took an extremely deep RNA-seq experiment of 1,095,239,309 mapped single end 100 base pair reads from a murine model of acute myeloid leukemia. We obtained 74,722,599 distinct reads, 173,485 observed RefSeq coding exons, and 402,196 unique junctions. The total number of observed exons and junctions was respectively 820,653,733 and 235,107,245. The coefficients of variation are 39.8, 5.1, and 8.9 for the read (Lib-Size), exonic (Exons), and slice junction (Junctions) populations. For the immune repertoire we examined the combined T-cell receptor β repertoire of 39 individuals of varying ages [4], using the inferred amino acid sequence to determine the repertoire (TCR). The combined data contained 16,704,121 TCR β species with 38,701,991 total observations and a CV of 66.8. To simulate metagenomic sampling experiments we took α -diversity source data from the MG-RAST database [31]. We examined two populations: soil microbiome of the Penang Mangrove forest (PenangMangrove, experiment ID 4510219.3) and the combined population of 14 marine sediment microbiome samples from the Gulf of Mexico following the Deepwater Horizon oil spill (GulfSediment, project ID

3023) These contained, respectively, 69,454 and 41,720 annotated species with 247,106,874 and 577,099,067 total observations and CVs of 11 and 11.1. All of the observed count frequencies are available in the Supplementary Materials.

For each population we examined three levels of sampling, varying across three orders of magnitude. We selected these levels based on existing experiments to approximately correspond to shallow, moderate, and deep sampling. For the RNA-seq experiments we sampled levels equivalent to 500,000 (500K), 5,000,000 (5M), and 50M mapped reads. In the metagenomic sampling experiments we sampled 10K, 100K, and 1M annotated species. Finally, in the T-Cell repertoire we sampled 100K, 1M, and 10M T-Cell β sequences (TCS). At each level we took 100 independent samples from the constructed samples.

We measured the average performance of the Chao estimator and the moment-based quadrature estimators with the median species richness estimate (\hat{S}). We also considered the variability of the estimates, quantified by the standard error (SE). There is a tradeoff to consider between the average estimate and variability. For example, one can use the observed number of species as an estimate of species richness. This simple estimator will have low variability but is typically a very poor estimator. To capture the bias-variance trade-off we used the root mean square error (RMSE). These results are summarized in Table 2.

The moment-based quadrature estimators were larger than the Chao estimator in all samples and in only 3 out of the 600 deepest samples was the MQE larger than the true species richness. The CE, M2QE, and M3QE always underestimated the true species richness, as did the MQE at the lower sampling depths. This method of simulation essentially constructs high, but finite, dimensional mixtures. In these cases, using just a few moments is not sufficient to estimate the species richness. The use of more moments, at least through our framework, can only increase the estimated species richness. But this naturally increases the variance and we observe increasing variability in the maximum number of moments used. The question that naturally follows is if the increase in the estimated species richness is worth the larger variability.

In all cases the MQE had the lowest RMSE, indicating that the increasing accuracy of using more points is worth the increase in variance. Sometimes the improvement was small, as was the case in the metagenomic samples, while in some cases the MQE was an order of magnitude larger than the CE, as was the case in the Exons population with a 500K read sample. Interestingly, for the metagenomic populations (GulfSediment & PenangMangrove) the MQE and M3QE were exactly equal in all but a few cases. This is most likely due to the fact that the MQE used more than 3 points extremely infrequently for the estimates in this case.

We noted the estimated species richness for the metagenomic populations is several times lower than the true species richness at the deepest level of sampling. Even when sampling at an order of magnitude larger (10M sequences from annotated species), the MQE still significantly underestimates the metagenomic species richness. This may indicate that larger experiments are needed to accurately quantify the rarest species in metagenomic sequencing experiments. On the other hand, it is clear that deep sampling or sequencing is required to capture transcriptional diversity, including both exonic and transcriptional diversity, and T-Cell repertoire diversity. It may be that our analysis ignores extremely rare sequences missing from the original population so that even deeper sampling is required in practice.

Population Sample Size	S	CE				M2QE				M3QE				MQE			
		\tilde{S}	SE	RMSE	\tilde{S}	$\hat{\sigma}_S$	RMSE	\tilde{S}	$\hat{\sigma}_S$	RMSE	\tilde{S}	$\hat{\sigma}_S$	RMSE	\tilde{S}	$\hat{\sigma}_S$	RMSE	RMSE
LibSize	7.47E7	CE				M2QE				M3QE				MQE			
500K reads	2.56E6	1.79E4	7.2E7	7.2E7	3.88E6	1.06E5	7.08E7	4.81E6	7.23E5	6.97E7	4.95E6	8.78E5	6.96E7	4.95E6	8.78E5	6.96E7	
5M reads	9.33E6	1.48E4	6.54E7	6.13E7	1.34E7	8.83E4	6.13E7	1.66E7	5.48E5	5.81E7	1.76E7	9.27E5	5.71E7	1.76E7	9.27E5	5.71E7	
50M reads	3.07E7	1.91E4	4.40E7	3.44E7	4.03E7	8.34E4	3.44E7	4.67E7	4.87E5	2.81E7	5.25E7	3.70E6	2.19E7	5.25E7	3.70E6	2.19E7	
Exons	1.73E5	CE				M2QE				M3QE				MQE			
500K reads	8.74E4	4.44E2	8.61E4	7.64E4	9.69E4	1.69E3	7.64E4	1.0052E5	3.60E3	7.28E4	1.0053E5	3.73E3	7.27E4	1.0053E5	3.73E3	7.27E4	
5M reads	1.23E5	3.13E2	5.02E4	4.38E4	1.30E5	1.20E3	4.38E4	1.326E5	2.6E3	4.06E4	1.327E5	3.16E3	4.04E4	1.327E5	3.16E3	4.04E4	
50M reads	1.51E5	2.94E2	2.22E4	1.73E4	1.56E5	1.11E3	1.73E4	1.578E5	2.12E3	1.552E4	1.579E5	2.18E3	1.545E4	1.579E5	2.18E3	1.545E4	
Junctions	4.02E5	CE				M2QE				M3QE				MQE			
500K reads	6.37E4	6.81E2	3.38E5	3.25E5	7.66E4	2.79E3	3.25E5	8.07E5	4.49E3	3.22E5	8.09E5	4.82E3	3.21E5	8.09E5	4.82E3	3.21E5	
5M reads	1.19E5	4.35E2	2.83E5	2.66E5	1.36E5	2.27E3	2.66E5	1.46E5	7.44E3	2.543E5	1.47E5	8.34E3	2.536E5	1.47E5	8.34E3	2.536E5	
50M reads	2.29E5	9.58E2	1.74E5	1.38E5	2.64E5	3.32E3	1.38E5	2.89E5	1.76E4	1.13E5	2.90E5	2.00E4	1.10E5	2.90E5	2.00E4	1.10E5	
TCR	1.67E7	CE				M2QE				M3QE				MQE			
100K TCS	1.55E6	3.28E4	1.52E7	1.42E7	2.43E6	2.11E5	1.42E7	2.68E6	3.03E5	1.401E7	2.69E6	3.05E5	1.400E7	2.69E6	3.05E5	1.400E7	
1M TCS	6.11E6	3.13E4	1.06E7	8.1E6	8.59E6	1.45E5	8.1E6	1.006E7	9.1E5	6.49E6	1.02E7	1.11E6	6.31E6	1.02E7	1.11E6	6.31E6	
10M TCS	1.40E7	1.47E4	2.70E6	1.21E6	1.55E7	4.31E4	1.21E6	1.60E7	1.26E5	6.40E5	1.62E7	1.98E5	5.25E5	1.62E7	1.98E5	5.25E5	
GulfSediment	4.17E4	CE				M2QE				M3QE				MQE			
10K	1.98E3	6.43E1	3.97E4	3.947E4	2.24E3	1.64E2	3.947E4	2.2534E3	1.7859E2	3.9452E4	2.2534E3	1.7859E2	3.9452E4	2.2534E3	1.7859E2	3.9452E4	
100K	3.89E3	1.28E2	3.78E4	3.68E4	4.74E3	6.65E2	3.68E4	4.8205E3	7.587E2	3.67258E4	4.8205E3	7.586E2	3.67252E4	4.8205E3	7.586E2	3.67252E4	
1M	9.10E3	2.25E2	3.26E4	3.04E4	1.11E4	1.20E3	3.04E4	1.1312E4	1.491E3	3.0092E4	1.1312E4	1.493E3	3.0089E4	1.1312E4	1.493E3	3.0089E4	
PenangMangrove	6.95E4	CE				M2QE				M3QE				MQE			
10K	1.95E3	4.81E1	6.75E4	6.732E4	2.11E3	1.47E2	6.732E4	2.1181E3	1.5531E2	6.7312E4	2.1181E3	1.5532E2	6.7312E4	2.1181E3	1.5532E2	6.7312E4	
100K	4.67E3	2.34E2	6.48E4	6.35E4	5.81E3	9.39E2	6.35E4	5.9588E3	1.076405E3	6.335329E4	5.9588E3	1.076406E3	6.335327E4	5.9588E3	1.076406E3	6.335327E4	
1M	1.16E4	3.29E2	5.78E4	5.33E4	1.55E4	2.45E3	5.33E4	1.602E4	3.195E3	5.2515E4	1.603E4	3.198E3	5.2509E4	1.603E4	3.198E3	5.2509E4	

Table 2: Performance of the Chao and moment-based quadrature estimators on numerically simulated populations taken from large observed sampling experiments. Bold indicates the highest median, lowest variance, and lowest RMSE estimators for a given population and sample size.

7 Discussion

The general moment-based framework for estimating species richness was developed over 50 years ago by Harris [18] and later explored by Chao [7], yet the practical application has been impeded by two primary issues: the lack of deep sampling experiments where higher moments can be accurately estimated, and efficient algorithms to compute the estimates. Inspired by the recent explosion of deep sampling experiments brought on by technological improvements, we presented algorithms specifically designed to calculate species richness in the moment-based framework of Harris to use the maximum amount of information possible from the observed species counts. We showed that our method improves upon the Chao estimator in the presence of heterogeneity and will default to the Chao estimator in the worst case.

There have been a multitude of new approaches in recent years for estimating species richness, including penalized non-parametric maximum likelihood estimation [39] and non-parametric non-linear regression of the count frequency ratios [40]. The moment-based quadrature estimator provides a new avenue of research into estimating species richness. Possible directions of future research include investigating algorithmic improvements for estimating the three term recurrence in the presence of error, a woefully under explored area of research that we briefly discussed in this paper. Such work to combat the ill-conditioning in the algorithm and improve the stability of the estimates will facilitate the use of more moments in the estimator. Our work suggests new conditions for smoothing count frequencies, a topic with a long history in the species richness problem [15, 16]. Rather than smoothing the count frequencies directly, one can smooth the moments through the moment Hankel matrices to ensure that the moments correspond to a true distribution, similar to the problem of finding the nearest covariance matrix [20], but with the added condition of constant off-diagonals.

The development of new sampling technology, particularly in high-throughput sequencing, is increasing our understanding of large and highly heterogeneous populations. The development of fast, efficient, and accurate algorithms will help further our understanding of these populations. The method we presented represents a significant improvement on previous non-parametric moment based estimators and allows for fast and accurate estimates of species richness. We have made the software available as part of the open source preseq package (<http://smithlabresearch.org/software/preseq/>).

Acknowledgements

The authors would like to thank Mike Waterman, Simon Tavaré, and Peter Calabrese for their advice on this paper. A special thanks goes to Sergey Lototsky for a simplified proof of Proposition 1. This work was funded by NIH grant R01 HG007650.

References

- [1] Piero Barone. On the universality of the distribution of the generalized eigenvalues of a pencil of Hankel random matrices. *Random Matrices: Theory and Applications*, 2(01), 2013.
- [2] Bernhard Beckermann and Emmanuel Bourreau. How to choose modified moments? *Journal of Computational and Applied Mathematics*, 98(1):81–98, 1998.
- [3] Dankmar Böhning. Some general comparative points on Chao’s and Zelterman’s estimators of the population size. *Scandinavian Journal of Statistics*, 37(2):221–236, 2010.

- [4] Olga V Britanova, Ekaterina V Putintseva, Mikhail Shugay, Ekaterina M Merzlyak, Maria A Turchaninova, Dmitriy B Staroverov, Dmitriy A Bolotin, Sergey Lukyanov, Ekaterina A Bogdanova, Ilgar Z Mamedov, et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *The Journal of Immunology*, 192(6):2689–2698, 2014.
- [5] John Bunge and M Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.
- [6] AD Carothers. Capture-recapture methods applied to a population with known parameters. *The Journal of Animal Ecology*, pages 125–146, 1973.
- [7] Anne Chao. *The Quadrature Method in Inference Problems Arising from the Generalized Multinomial Distribution*. PhD thesis, University of Wisconsin-Madison, 1977.
- [8] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pages 265–270, 1984.
- [9] Anne Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791, 1987.
- [10] Didier Dacunha-Castelle and Elisabeth Gassiat. The estimation of the order of a mixture model. *Bernoulli*, pages 279–299, 1997.
- [11] Philip J Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Courier Dover Publications, 2007.
- [12] Walter Gautschi. On generating orthogonal polynomials. *SIAM Journal on Scientific and Statistical Computing*, 3(3):289–317, 1982.
- [13] Walter Gautschi. *Orthogonal Polynomials: Computation and Approximation, Numerical Mathematics and Scientific Computation Series*. Oxford University Press, Oxford, 2004.
- [14] Gene H Golub and John H Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969.
- [15] IJ Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.
- [16] IJ Good. The use of divergent series in statistical problems. *Journal of Statistical Computation and Simulation*, 34(2-3):164–164, 1990.
- [17] Richard Hamming. *Numerical Methods for Scientists and Engineers*. Dover Publications, 1987.
- [18] Bernard Harris. Determining bounds on integrals with applications to cataloging problems. *The Annals of Mathematical Statistics*, 30(2):521–548, 1959.
- [19] Ian R Harris. The estimated frequency of zero for a mixed Poisson distribution. *Statistics & probability letters*, 12(5):371–372, 1991.
- [20] Nicholas J Higham. Computing the nearest correlation matrix: a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.
- [21] S Karlin and LS Shapley. *Geometry of moment spaces*. American Mathematical Society, 1953.

- [22] Samuel Karlin and William J Studden. *Tchebycheff Systems: With Applications in Analysis and Statistics*, volume 376. Interscience Publishers New York, 1966.
- [23] Dirk P Laurie. Computation of Gauss-type quadrature formulas. *Journal of Computational and Applied Mathematics*, 127(1):201–217, 2001.
- [24] Bruce G Lindsay. Moment matrices: applications in mixtures. *The Annals of Statistics*, pages 722–740, 1989.
- [25] Bruce G Lindsay. On the determinants of moment matrices. *The Annals of Statistics*, pages 711–721, 1989.
- [26] Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163. JSTOR, 1995.
- [27] William A Link. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4):1123–1130, 2003.
- [28] Chang Xuan Mao. Lower bounds to the population size when capture probabilities vary over individuals. *Australian & New Zealand Journal of Statistics*, 50(2):125–134, 2008.
- [29] Chang Xuan Mao and Bruce G Lindsay. Estimating the number of classes. *The Annals of Statistics*, pages 917–930, 2007.
- [30] Chang Xuan Mao and Na You. On comparison of mixture models for closed population capture–recapture studies. *Biometrics*, 65(2):547–553, 2009.
- [31] Folker Meyer, Daniel Paarmann, Mark D’Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.
- [32] Beresford N Parlett. *The Symmetric Eigenvalue Problem*, volume 7. SIAM, 1980.
- [33] RA Sack and AF Donovan. An algorithm for Gaussian quadrature given modified moments. *Numerische Mathematik*, 18(5):465–478, 1971.
- [34] Barry Simon. The classical moment problem as a self-adjoint finite difference operator. *Advances in Mathematics*, 137(1):82–203, 1998.
- [35] Gabor Szegő. *Orthogonal Polynomials*, volume 23. Amer Mathematical Society, 1975.
- [36] Pafnuty Tchébychev. *Sur l’interpolation par la méthode des moindres carrés*. Eggers et Comp., 1859.
- [37] Evgenij E Tyrtyshnikov. How bad are Hankel matrices? *Numerische Mathematik*, 67(2):261–269, 1994.
- [38] Ji-Ping Wang and Bruce G Lindsay. An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology*, 5(1):30–45, 2008.
- [39] Ji-Ping Z Wang and Bruce G Lindsay. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, 100(471):942–959, 2005.
- [40] Amy Willis and John Bunge. Estimating diversity via frequency ratios. *Biometrics*, 2015.

A

A.1 Proof of Proposition 1

Proof. Recall the polynomial given in equation 8,

$$L_i^{(r,\phi)}(x) = \sum_{l=0}^i \binom{i+r}{i-l} \frac{i!}{l!} \phi^{l-i} (-1)^{i+l} x^l.$$

We will show that these polynomials are the only system of polynomials orthogonal under the measure $d\nu(x) = x^r e^{-x\phi}$ for $r > 0$ and $\phi > 0$. To this end first note that the polynomial can be rewritten in the Rodrigues' formula

$$L_i^{(r,\phi)}(x) = (-1)^i \phi^i (x^{-r} e^{x\phi}) \left(\frac{d}{dx} \right)^i (x^{i+r} e^{-x\phi}). \quad (9)$$

Consider the classical associated Laguerre polynomials [35], orthogonal under the measure $d\omega(x) = x^r e^{-x} dx$ and given by

$$\begin{aligned} L_i^{(r)}(x) &= \sum_{l=0}^i \binom{i+r}{i-l} \frac{i!}{l!} (-1)^{i+l} x^l \\ &= (-1)^i (x^{-r} e^x) \left(\frac{d}{dx} \right)^i (x^{i+r} e^{-x}). \end{aligned} \quad (10)$$

By orthogonality the following holds for all positive integers $i \neq j$,

$$\int_0^\infty L_i^{(r)}(x) L_j^{(r)}(x) d\omega(x) = 0.$$

The left hand side above can be expanded as

$$\int_0^\infty (-1)^i (x^r e^x) \left(\frac{d}{dx} \right)^i (x^{i+r} e^{-x}) (-1)^j (x^{-r} e^x) \left(\frac{d}{dx} \right)^j (x^{j+r} e^{-x}) x^r e^{-x} dx = 0. \quad (11)$$

Consider the substitution $x = \phi y$ with $dx = \phi dy$. Then the left hand side of equation (11) above is equal to

$$\begin{aligned} &\int_0^\infty (-1)^i ((\phi y)^{-r} e^{\phi y}) \left(\frac{d}{\phi dy} \right)^i ((\phi y)^{i+r} e^{-\phi y}) \\ &\quad \cdot (-1)^j ((\phi y)^{-r} e^{\phi y}) \left(\frac{d}{\phi dy} \right)^j ((\phi y)^{j+r} e^{-\phi y}) (\phi y)^r e^{-\phi y} \phi dy \\ &= \phi^{r+1-i-j} \int_0^\infty (-1)^i \phi^i (y^{-r} e^{\phi y}) \left(\frac{d}{dy} \right)^i (y^{i+r} e^{-\phi y}) (-1)^j \phi^j (y^{-r} e^{\phi y}) \left(\frac{d}{dy} \right)^j (y^{j+r} e^{-\phi y}) y^{-r} e^{-\phi y} dy \\ &= \phi^{r+1-i-j} \int_0^\infty L_i^{(r,\phi)}(y) L_j^{(r,\phi)}(x) d\nu(y). \end{aligned} \quad (12)$$

By equation (11), the above integral in equation (12) must equal 0. It immediately follows that for all $i \neq j$ we have

$$\int_0^\infty L_i^{(r,\phi)}(y) L_j^{(r,\phi)}(x) d\nu(y) = 0.$$

Note that trivially $\int_0^\infty (L_i^{(r,\phi)}(y))^2 d\nu(y) > 0$ for all non-negative integers n . Therefore the system of monic polynomials $\{L_i^{(r,\phi)}(x)\}_{i \geq 0}$ is orthogonal under the measure $d\nu(x) = x^r e^{-x\phi} dx$.

Since the monic polynomials can be constructed by the Gram-Schmidt algorithm, uniqueness follows immediately from the positive definiteness of the measure $d\nu$. \square

A.2 Proof of Proposition 2

Proof. Since $\{L_i^{(r,\phi)}(x)\}_{i \geq 0}$ is a system of monic polynomials orthogonal under the measure $d\nu$, $\{L_i^{(r,\phi)}(x)\}_{i \geq 0}$ must satisfy the three term recurrence of equation (7) for some $\alpha_i > 0$ and $\beta_i > 0$. Therefore, the following holds for all $i \geq 0$,

$$L_{i+1}^{(r,\phi)}(x) = (x - \alpha_i)L_i^{(r,\phi)}(x) - \beta_i L_{i-1}^{(r,\phi)}(x). \quad (13)$$

Substituting $L_i^{(r,\phi)}(x)$ defined in equation (8) results in the following equation

$$\begin{aligned} & \sum_{l=0}^{i+1} \binom{i+1+r}{i+1-l} \frac{(i+1)!}{l!} \phi^{l-i-1} (-1)^{i+1+l} x^l \\ &= (x - \alpha_i) \sum_{l=0}^i \binom{i+r}{i-l} \frac{i!}{l!} \phi^{l-i} (-1)^{i+l} x^l \\ & \quad - \beta_i \sum_{l=0}^{i-1} \binom{i-1+r}{i-l} \frac{(i-1)!}{l!} \phi^{l-i+1} (-1)^{i-1+l} x^l. \end{aligned} \quad (14)$$

Since equation (14) must hold for all x , the reduced coefficients of both sides must be equal. We can group terms on both sides of the equation according to the degree of x to solve for α_i and β_i .

These polynomials are monic, so the highest order term is x^{i+1} . The term of second highest order, as written on both sides of the equation, simplifies to

$$\frac{\Gamma(i+r+2)(i+1)!}{\Gamma(i+1+r)i!} \phi^{-1} (-1)^i x^i = \frac{\Gamma(i+r+1)i!}{\Gamma(i+r)(i-1)!} \phi^{-1} (-1)^{-1} x^i - \alpha_i x^i.$$

Further simplifying the coefficients of x^i results in

$$(i+r+1)(i+1) = (i+r)i + \alpha_i \phi,$$

providing

$$\alpha_i = (2i+r+1)/\phi.$$

Now consider the lowest order terms. The constants written on both sides of equation (14) are

$$\frac{\Gamma(i+r+2)}{\Gamma(r+1)} \phi^{-i-1} (-1)^{i+1} = -\alpha_i \frac{\Gamma(i+r+1)}{\Gamma(r+1)} \phi^{-i} (-1)^i - \beta_i \frac{\Gamma(i+r)}{\Gamma(r+1)} \phi^{-i+1} (-1)^{i-1}$$

After simplification, and substitution with α_i above, we arrive at

$$\beta_i = (i^2 + ir)/\phi^2.$$

The above calculated α_i and β_i , obtained using the second highest and lowest order coefficients, can be used to verify the equality of equation (14) for all other coefficients. \square